Journal of Informatics and Web Engineering

Vol. 3 No. 3 (October 2024) eISSN: 2821-370X

Comparative Evaluation of Machine Learning Models for Mobile Phone Price Prediction: Assessing Accuracy, Robustness, and Generalization Performance

Saima Anwar Lashari[1], Muhammad Muntazir Khan[2], Abdullah Khan[3*], Sana Salahuddin[4], Muhammad Noman Atta[5]

[1]College of Computing and Informatics, Saudi Electronic University, 4552 Prince Mohammed Ibn Salman Ibn Abdulaziz Rd, 6867, Ar Rabi, Riyadh 13316, Saudi Arabia.

[2,3,4,5]Institute of Computer Sciences and Information Technology, The University of Agriculture Peshawar, Peshawar, Khyber Pakhtunkhwa, Pakistan.

**corresponding author: (abdullah_khan@aup.edu.pk; ORCiD: 0000-0003-1718-7038)*

Abstract - These days, mobile phones are the most commonly purchased goods. Thousands of new models with improved features, designs, and specifications are released yearly. An autonomous mobile price prediction system is required to assist customers in determining whether or not they can afford these devices. Many machine learning models exhibit varying performance degrees based on their architecture and learning properties. Ten widely used classifiers were assessed in this study: Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Decision Tree (DT), Naïve Bayes (NB), Linear Discriminant Analysis (LDA), AdaBoost, and Light Gradient Boosting (LGB). The F1-score, recall, accuracy, and precision of these models were evaluated. According to the findings, the results indicated that LR, with its use of the Elastic Net parameter, outperformed the others with 96% accuracy, 97% precision, 94% recall, and 96% F1-score. Other models like XGBoost, LGB, and SVM also showed strong performance, whereas KNN had the poorest performance. The study highlights the importance of selecting the appropriate model for accurate mobile price prediction. Among all the machine learning used in this paper, the LR classifier outperforms the other state-of-the-art models because of the elastic Net parameter used for mobile phone price prediction.

Keywords— Logistic Regression, Random Forest, K-Nearest Neighbor, Support Vector Machine, Extreme Gradient Boosting, Decision Tree, Naïve Bayes, Linear Discriminant Analysis, AdaBoost, Light Gradient Boosting

Received: 13 June 2024; Accepted: 05 August 2024; Published: 16 October 2024

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2024.3.3.9
© Universiti Telekom Sdn Bhd.
Published by MMU Press. URL: journals.mmupress.com/jiwe

1. INTRODUCTION

Price is usually a significant determining element in the decision to purchase a product and in the buyer's thinking when deciding "what the worth is and whether it is good to buy in this range". Many variables and factors are taken into account when a product is introduced to the market, and this is especially true for mobile devices, where many features and specifications, such as memory, are taken into account. The impact of the price may also impact the level of market competition. Mobile devices have numerous specifications and features, including camera, video, CPU quality, and material quality. There are several restrictions regarding pricing considerations because the product needs to be affordable and accessible when taken as a whole [1]. The process of predicting the eventual price of a mobile phone using machine learning and statistical approaches is known as mobile phone price prediction. This can be achieved by looking back at past sales, prices, and market patterns for mobile phones and using that data to estimate future prices [2]. The same process can be used to determine the true cost of any goods, including automobiles, bikes, generators, motors, food, medicine, and others [3]. Different methods, such as linear regression, time series analysis, decision trees, and neural networks, can be used to predict mobile phone prices. Utilizing pertinent data, such as economic indicators and consumer mood, as well as front-line machine learning algorithms, with deep learning or reinforcement learning, will increase the prediction's accuracy [4]. A lot of mobile phones of different varieties are produced by various brands rapidly. So, the cost of mobile based on its attributes is efficient. Machine learning algorithms can be used to perform said task [5]. Price is the most effective commercial and marketing attribute. The pricing of goods is the very first query in industries. The best learning methods, such as unsupervised and supervised learning like Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) are provided by machine learning for prediction purposes [6]. Mobile phones are the most used item nowadays. Different phone brands produce various phone types with new and advanced specifications every year. The availability of various mobile phone types confuses customers when choosing a specific phone. As the customer always wants to purchase the best feature mobile phone with a small amount of money, they need to know the price of various phones with desired features. So, an automated system is needed to predict the price of a mobile phone by its features. Therefore, in the need to find out cost prediction for mobile phone with high accuracy, this research study compares different machine learning models for mobile cost prediction. Ten different machine learning models analyze the data from an open-source Internet repository. This paper assessed each classification method's precision, recall, accuracy, and F1-measure by examining the "Mobile phone price prediction" dataset collected from Kaggle. This paper consists of the following contributions:

- Analyzing the dataset with ten different machine learning models is necessary for better results. K-Nearest Neighbor (KNN), Naïve Bayes (NB), DT, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGB), Linear Discriminant Analysis (LDA), SVM, RF and LR are evaluated for mobile cost prediction.
- Achieving high accuracy to ensure the best prediction results to help business individuals and customers improve financial management, better resource allocation, better customer management, increased profitability, better decision making, and improved data management.

The paper has four sections. The previous work is explained in Section 2. While the training model and proposed methodology is described in Section 3, the findings and analyses are offered in Section 4, and the conclusions and recommendations are obtainable in Section 5.

2. LITERATURE REVIEW

Various researchers used machine learning and deep learning models for mobile cost prediction. Some of these researchers and their research work are described as follows. Some tools, like WEKA, were used for the above task. Similarly, the author in [1] proposed a methodology for mobile price prediction. The author used K-Mean clustering along with above method for mobile cost prediction. The dataset is collected from Google. The author described that the said method achieved better results in the field of mobile cost prediction. Furthermore, [2] used the same (our proposed) dataset for mobile cost prediction. This paper trained five machine learning algorithms for the proposed task, like DT, RF, LR, KNN, LDA, and NB. From the overall process, it is claimed that LR achieved better results than other models for the above task. In [3], the author proposed multivariate regression by using WEKA tool for mobile cost prediction. To check the model's performance, the dataset is collected from the openly available repository www.GSMArena.com. The overall simulation shows that the proposed model achieved high accuracy for the mobile cost prediction. A self-collected data set is used for mobile cost prediction by [4]. To check the effectiveness of collected data, the author used three machine learning models: RF, SVM, and LR. The simulation results show that the LR achieved better results in terms of accuracy, recall, precision and F-Measure compared to other models. Another paper in [5] used WEKA tool for mobile cost prediction. The proposed study has used dimensionality reduction, feature selection, feature extraction, forward selection and backward selection as preprocessing steps to ensure high accuracy. Furthermore ZeroR, NBand J48 were selected as machine learning classifiers for mobile cost prediction. The simulation results show that NBachieved better results from other used models.

The comparison of KNN and LR is proposed by [6] for mobile cost prediction. To check the performance of the proposed models, the author has collected data randomly. Some preprocessing steps have been applied to the data set in the proposed study to ensure high accuracy in mobile cost prediction. The author used machine learning models such as LR and NB. The simulation results proposed that LR achieved the best results in terms of accuracy, precision and recall. The same models have been used by [7], Kuiper used a multivariate regression model to predict General Motors vehicle prices for 2005. To assess the efficacy of the proposed models, the author collected data from www.pakwheels.com, an easily accessible website. An alternative investigator uses the SVM model [8] for the initiative on leasing car prices. The author employed the previously described technique to predict the cost of leased cars. This study found that the SVM methodology is far more accurate and successful in predicting prices than other approaches, such as multiple LR, when a large data set is provided.Additionally, the study showed how SVM avoids over- and under-fitting issues and more skillfully handles high-dimensional data. To identify important characteristics for SVM, the author employed a genetic approach. However, the approach failed to demonstrate why SVM is superior to straightforward multiple regression in terms of variance and mean standard deviation. Similarly another paper in [9] estimates the cost of used automobiles in Mauritius. The author downloads the data set from Google to evaluate the effectiveness of different machine learning models. The author used various techniques to estimate the prices, including numerous LR, KNN, DT, and NB. It was learned during the investigation that the most well-liked algorithms like KNN and NB, has comparable results. However, the DT achieved the best results in terms of accuracy, precision, and recall.

Furthermore, in [10], the Elman Neural Network (CSENN) is combined with the recommended Chicken-Swarm-Optimization (CSO) to maximize the learning weights of the Elman Neural Network. The IRIS and 7-bit parity classification datasets were used to evaluate the proposed models' effectiveness. The outcomes are contrasted with those of genetic algorithm neural networks (GANN), artificial bee colony back propagation (ABCBP), and BPNN. The simulation's findings demonstrate that the CSENN approach outperforms other traditional algorithms in terms of accuracy and mean square error. The proposed CSO is combined with the existing Elman CSENN to optimize the learning parameters of the ENN. Two classification datasets, namely 7-bit parity and IRIS, were employed to evaluate the effectiveness of the proposed models. The outcomes are contrasted with those of genetic algorithm neural networks (GANN), artificial bee colony back propagation (ABCBP), and BPNN. The simulation's findings demonstrate that the CSENN approach outperforms other traditional algorithms with respect of accuracy as well as mean square error. The previous work discussed above states that different researchers have used machine learning models in a wide range of work. Various research studies have been conducted for mobile price prediction using machine learning and deep learning models. The proposed study is conducted as a performance assessment of different machine learning models for the classification and prediction task.

3. TRAINING ALGORITHMS

The recommended research study suggested various machine learning models for mobile phone price prediction. Some of the methods used with high efficiency are discussed as follows.

3.1 Support Vector Machine (SVM)

The SVM, created by Vapnik in 1995, is a supervised learning method with extensive applications in statistical regression analysis and classification [17]. The distinction between the hyperplanes with the largest negative and positive sample gaps in the feature space (FS) is what the initial SVM model set out to identify. When the insensitive loss function (-ILF) was first introduced, SVM was only utilized to address classification problems. Regression-related tasks, such as linear or non-linear, were carried out using it. The SVM techniques will be used in this work as a regression analyzer. As a result, this part will present the SVM technique and explain how to use it to overcome regression issues. $(x1,\ y1),\ (x2,\ y2)(xn,\ yn)$ is a set of training data, $Dk\ =\ (xk,\ yk)$ is the total quantity of training data, and n is the kth train sample. The SVM regression seeks to bring all sample points closer to the hyperplane to minimise the overall deviation between the sample points and the hyperplane. If there is a linear relationship between these sample points, Equation (1) shows the linear regression function.

$$\boldsymbol{f(x) = w.x + e} \tag{1}$$

A non-linear function (x) in non-linear regression problems maps each sample point to a high-dimensional FS. In this way, the original space's non-linear regression is transformed into a linear problem, and the linear regression analysis is then carried out using the high-dimensional FS. Consequently, Equation (2) shows the representation of the SVM regression's decision function is possible [10].

$$\boldsymbol{f(x) = w.\phi(x) + \mathrm{e}(2)} \tag{2}$$

3.2 Extreme Gradient Boosting (XGBoost)

XGBoost is a specific application of gradient boosting techniques primarily used to address challenging classification and regression issues. The XGBoost concept was employed in numerous types of research to calculate the vulnerability to natural hazards. The use of the XGBoost model in the research projects mentioned above yields extremely accurate findings for GEOCARTO International 9. One advantage of XGBoost is that it can shorten processing time by earning the ideal number of boosting iterations in a single run. Additionally, it is commonly known that by reducing overfitting, the XGBoost can improve modeling accuracy, which was another reason for the verdict to use this model on behalf of this study[11]. A special R studio code was used to utilize the XGBoost model and determine the flash-flood vulnerability. The functioning of the XGBoost model is described by Equation (3).

$$\boldsymbol{g(bt) = \sum_{i=1}^{n} R(pi, pi^{(m-1)} + bt(yi)\ ) +\ \Omega(bt) + c} \tag{3}$$

Where i stand for the "$\boldsymbol{ith}$ number of samples," "$\boldsymbol{pi}$" is the "m 1th model's projected value for the "$\boldsymbol{ith}$ sample,""$\boldsymbol{bt}$" stands for the "newly added model," "$\boldsymbol{X}$" is the "regular value," "$\boldsymbol{C}$" is the "constant value," and "R" stands for the "model error."

3.3 Logistic Regression (LR)

Binary classification is a task whereas the output variable only has two possible values, commonly expressed as 0 and 1. LR is a type of statistical model used for this task. The chance of the output variable attaining 1 due to the input factors is modelled using LR. The likelihood that the output variable will be 1 is modelled using the logistic and sigmoid functions. An S-shaped curve, the sigmoid function, transfers any real integer to the range [0, 1]. Equation (4) defines the LR model.

$$\boldsymbol{logit(y) = 1n(\frac{p}{1-p})}=\mathrm{ɑ+ß}x \tag{4}$$

Where p is the likelihood that the desired outcome will occur and x is the variable that explains the outcome. The variables for the LR are. This is how the fundamental logistic model looks [12].

4. RESEARCH METHODOLOGY

The research methodology for this proposed study involves several key steps. Firstly, publicly available datasets is collected and preprocessed to ensure data consistency and quality. Next, a machine learning [14] framework are designed and implemented to predicate mobile prices. Different data preprocessing techniques are implemented using python library. The learning of the proposed framework are measured using standard metrics, like accuracy, specificity, sensitivity, and confusion metrics. Comparative investigations are conducted to evaluate the proposed framework against current state-of-the-art methods. Figure 1 gives the research methodology flow diagram of this paper.

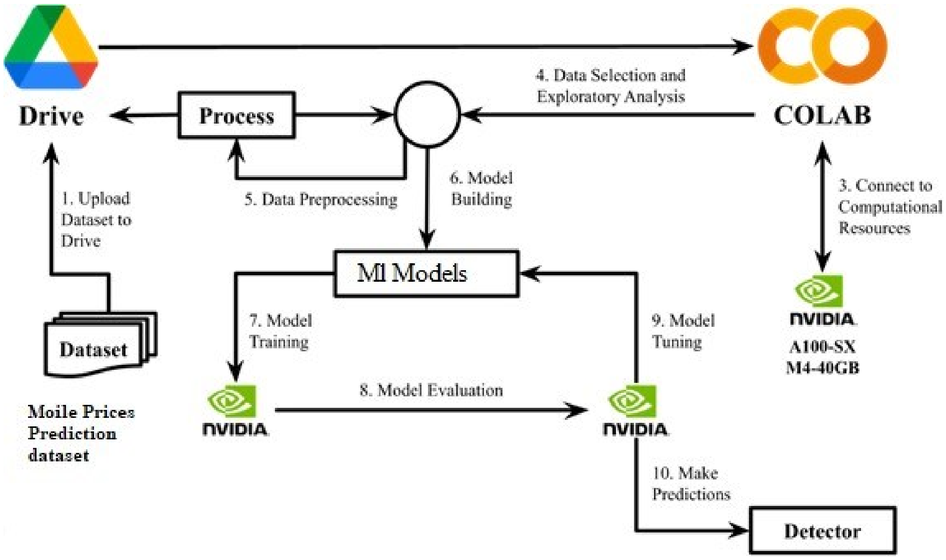
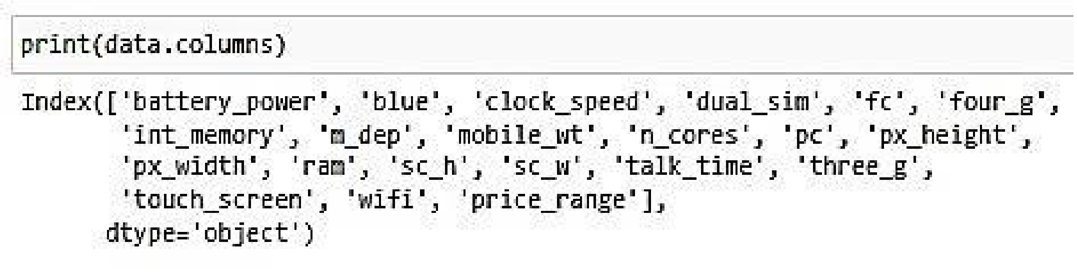


Figure 1. Proposed Research Methodology Flow Diagram

4.1 Data Collection

Dataset used in the proposed study is collected from the online repository Kaggle. The dataset is named "Mobile_price_predictia" and the link for the dataset is https://www.kaggle.com/code/amankumar1007/mobile-price-predictiona/data. Mobile features and data are gathered in many ways; android or another type of mobile device, screen size, RAM, camera pixel count, length and thickness of the device and battery life are all recorded. The data heads/ columns and records in the proposed dataset are elaborated in Figure 2 and Table 1.



```
print(data.columns)

Index(['battery_power', 'blue', 'clock_speed', 'dual_sim', 'fc', 'four_g',
       'int_memory', 'm_dep', 'mobile_wt', 'n_cores', 'pc', 'px_height',
       'px_width', 'ram', 'sc_h', 'sc_w', 'talk_time', 'three_g',
       'touch_screen', 'wifi', 'price_range'],
      dtype='object')
```

Figure 2. Description of Dataset

Table 1. Review of Attribute Contained In the Proposed Dataset

S.NO	Battery_power	Blue	Clock_speed	Dual_Sim	Fc	Four_g	Int_memory	…..	price
0	842	0	2.2	0	1	0	7	…..	1
1	1021	1	0.5	1	0	1	53	…..	
2	563	1	0.5	1	2	1	41	…..	2
3	615	1	2.5	0	0	0	10	…..	2
4	1821	1	1.2	1	13	1	44	…..	2
5	1859	0	0.5	1	3	1	44	…..	1
6	1821	0	1.7	0	4	1	10	…..	1
7	1954	0	0.5	0	0	0	53	…..	3

4.2 Data Analysis

Reviewing and evaluating data carefully to draw important conclusions is the data analysis procedure, form judgments, and support decision-making. A data analysis is a condensed and targeted analysis of particular data or a particular feature of data to respond to a particular query or realize a specific objective. The acquired data is examined, and the minimum and maximum values are taken into account along with the count and mean value [13]. The pricing value is segregated for greater accuracy into training data sets and instance data sets, and it is classified into other categories as indicated below. In machine learning, "correlation" describes the connection between a pair of variables and how changes in a single variable affect another (s). It explains the nature, direction, and intensity of the link between the variables. Correlation is frequently used in machine learning to find variables that are significant predictors inside a model as well as to identify and fix problems like multi co-linearity.

4.3 Data visualization

Data visualization, which includes charts, graphs, maps, info graphics and so on, is the depiction of information or data in some kind of a visual or pictorial format. It is an essential part of data analysis and aids in conveying complex information in a way that is simple to understand. With the help of data visualization, it is possible to take huge and intricate data sets and show them in a simple form, making it possible to see patterns, trends, and insights right away. Visualizing data makes finding relationships, identifying outliers, and making defensible decisions based on the information provided simpler.

Figure 3 shows the dataset plotted. This graph consists of different attributes available in dataset. It also includes a visual display of values exist in various columns or records of dataset. There will be more complexity and difficulty in visualizing the data if there are more restrictions, features, and variables. Therefore, if the features are connected, we could employ complexity-reducing methods. By removing unneeded characteristics and keeping only those that provide the most precise information, feature selection is employed to choose a specific dimension. By using forward and backward selection, the dataset is extracted to locate the specified dataset and features. When selecting forward, we begin with no features and then increase the critical features that will reveal the critical information. In backward selection, we just maintain the key traits and discard any features that don't provide any information.

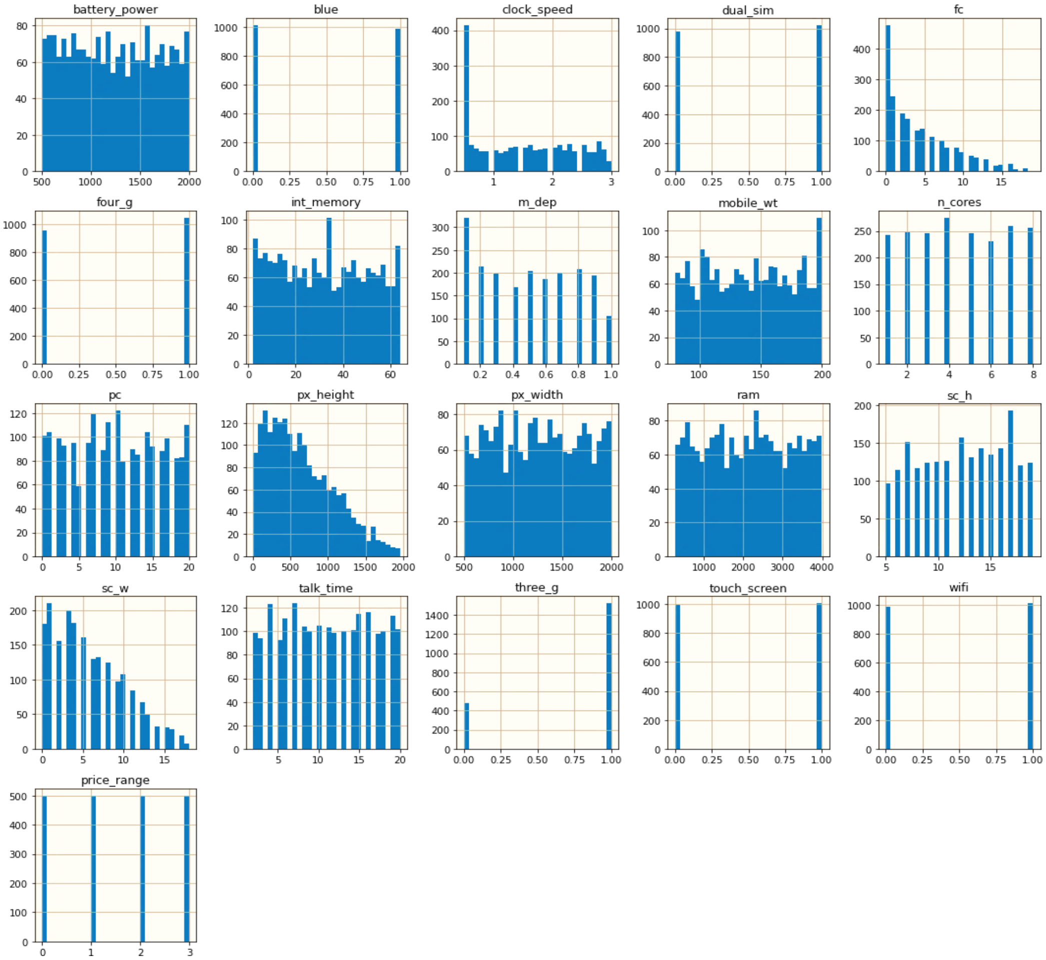


Figure 3. Visualization of Dataset

4.4 Proposed Model

While LR has achieved the best results as compared to other machine learning models, LR performs well in regression problems rather than classification. This study suggested a parameter tuning like elastic Net. Elastic Net can be achieved by combining lasso regression (L1) and ridge regression (L2) regularization. Elastic regularization performs very well in case of correlated data. As elastic regularization can be evaluated by root mean square error (RMSE) and mean absolute error (MAE), the performance evaluation uses matrices like accuracy, precision and recall by discretising the output. So, this study suggested LR as proposed model. So, the following is a suggested pseudocode for the proposed model.

Pseudo Code

```
Input data:        Give input data to the training models

Output Results:  Generate mobile prices predication

Start
Step 1: #input data
        X = input data matrix of shape (n_samples, n_features)
        y = target vector of shape (n_samples,)
```

```
Step 2: # Initialize model parameters
        Theta = array of zeros with shape (n_features + 1,)
        def sigmoid(z):
        Return 1 / (1 + exp(-z))
        X = add column of ones to X
        for i in range(n_iterations):
        h = sigmoid (X.dot(theta))
Step 3: # Calculate the cost function and gradient
        Cost = (-y.dot(log(h)) - (1-y).dot(log(1-h))) / n_samples
        Gradient = X.T.dot(h - y) / n_samples
Step 4: # Update the parameters
        Theta = theta - learning_rate * gradient
Step 5: Make final prediction
        y_pred = sigmoid (X.dot(theta))
Step 6: Evaluate the Model
        Accuracy = calculate accuracy of predictions y_pred against y
Step 7: End
```

4.5 Evaluation Parameters

Various parameters are used in this paper to check the model's performance. The efficiency of the proposed study is to compare the results of KNN, NB, LDA, DT, XGBoost, LGB, AdaBoost RF, LR, and SVM. Additional analyses and verifications of the performance of the suggested models are done using the following constraints.

4.5.1 Recall

Recall is the total number of positive values successfully detected and added together, and recall is computed by dividing that number by the total number of true positive and false negative values (see Equation (5)). "True Positive Rate" measurements refer to positive components that can be precisely detected. Cases with a high recall rate had the correct results.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

4.5.2 Precision

By dividing the total of positive results with false positives, precision is obtained by the total number for correctly detected values (see Equation (6)).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

4.5.3 F-Measure

To calculate F-Measure, recall and precision are utilized, as shown in Equation (7).

$$\text{F-Measure} = \frac{2*Recall*Precision}{Recall+Precision} \tag{7}$$

4.5.4 Accuracy

Accuracy reveals how confidently the model can distinguish between negative and positive classifications. It is calculable as in Equation (8).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

Here, TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively. This research uses a mobile cost prediction dataset to assess the efficiency of the proposed machine learning classifiers. The effectiveness of the suggested models is evaluated using accuracy, f-measure, precision, and recall. The models employed both testing and training data. An 80:20 split is used to divide the dataset into testing and training groups.

5. RESULT AND DISCUSSION

For this study, an Intel Core i7 PC running Windows 10 with 16 GB of RAM and a 2.0 GHz processor was employed. All datasets were utilized to test and train the algorithms using the Keras Python library. Ten distinct models, KNN, NB, LDA, DT, LR, XGBOOST, LGB, AdaBoost, RF and SVM are trained to provide the findings. The comparisons of several algorithms in terms of accuracy, F-Measure, recall, and precision were compiled.

5.1 Results Evaluation of Used Machine Learning Models

The performance of the used models, including KNN, NB, LDA, DT, SVM, RF, LR, XGBoost, AdaBoost and LGB are discussed in this section. From Table 2, boosting algorithms such as XGBoost, LGB, and linear discriminant analysis have performed superior since they produced more accurate findings. The accuracy of the XGBoost was 91%, the LGB was 90%, the LDA was 95%, and the performance of the KNN was inferior as it achieved 54% accuracy. Additionally, NB demonstrated a contented 82% accuracy, DT achieved an accuracy of 75%, and AdaBoost had 81% of accuracy. Furthermore, SVM achieved 88% accuracy, RF achieved 88%, while the LR classifier performed at 96% accuracy. So, based on the description above, the LR model exceeded the outcomes when compared to other models.

The above-mentioned LR model has attained a commendable performance of 96.33% accuracy in terms of Accuracy, Recall, Precision, and F-Measure. Table 2 and Figure 4 summarize the overall performance of various models, including KNN, NB, LDA, DT, SVM, RF, LR, XGBoost, AdaBoost and LGB. The XGBoost achieved 91% of accuracy, further, it achieved precision at 0.64%, recall at 0.58% and F-Measure at 0.60%. The accuracy of LGB is 0.90%, with precision of 0.93%, Recall of 0.95% and F-Measure of 92.68%. Furthermore, LDA achieved an accuracy of 95%, with precision of 0.96%, recall of 0.98% and F-Measure of 0.83%. After that, the performance of the KNN was poor as it achieved 54% accuracy, precision of 0.64%, recall of 0.58% and F-Measure of 0.60%. Additionally, NB demonstrated a contented 82% accuracy, precision of 0.89%, recall of 0.90% and F-Measure of 0.90%.

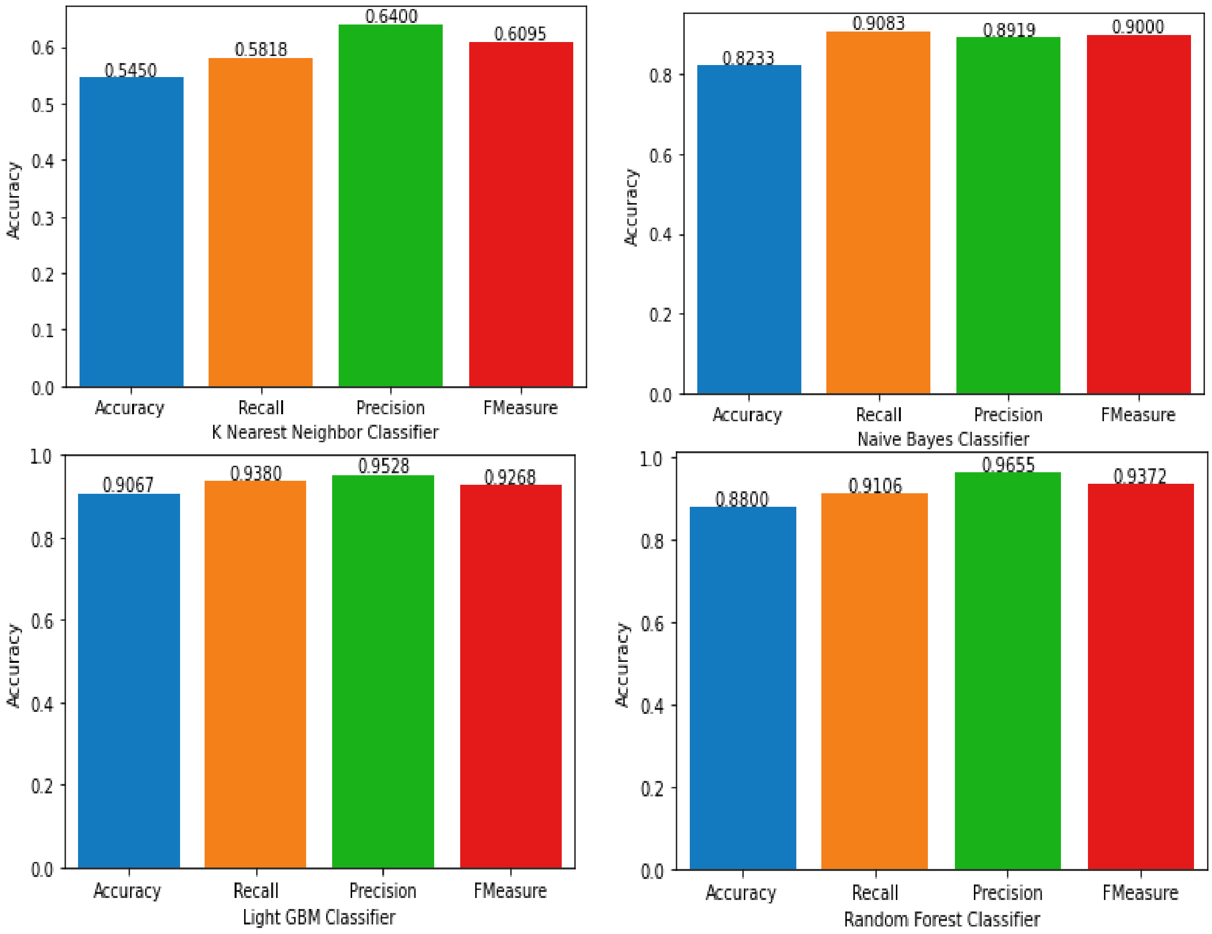
DT achieved accuracy of 75%, precision of 0.84%, recall of 0.78% and F-Measure of 0.81%. Also, AdaBoost has 81% of accuracy, precision of 0.77%, recall rate of 0.85% and F-Measure of 0.79%. Further, SVM achieved 88% of accuracy, precision of 0.95%, recall of 0.92% and F-Measure of 0.89%. Again, RF achieved 88% of accuracy, precision of 0.96%, and recall of 0.91% and F-Measure of 0.93%, while the LR classifier performed 96% of accuracy, precision of 0.97%, recall of 0.94% and F-Measure of 0.96%. Table 2 shows the performance of various models for mobile cost prediction in tabular form, while Figure 4 shows the graphical representation of performance evaluation of various models.

Table 2. Overall Performance Evaluation of The Used Models

Model	Accuracy	Precision	Recall	F-Measure
KNN	54.50	64.00	58.18	60.95
Naïve Bayes	82.33	89.18	90.83	90.00
DT	75.17	84.35	78.23	81.17

XGBoost	91.33	96.80	94.53	86.96
LGB	90.67	93.80	95.28	92.68
AdaBoost	81.67	77.31	85.19	79.47
LDA	95.00	96.95	98.45	83.74
SVM	89.83	95.90	92.86	89.25
RF	88.00	96.55	91.06	93.72
LR	96.33	97.66	94.70	96.15

Figure 4 visualizes the performance achieved by various models for mobile cost prediction. The graph shows that the XGBoost achieved 91%, with precision of 0.64%, recall rate of 0.58% and F-Measure of 0.60%. The LGB has an accuracy of 0.90%, with precision of 0.93%, Recall of 0.95% and F-Measure of 92.68%. Further, LDA achieved the accuracy of 95%, with a precision of 0.96%, recall of 0.98% and F-Measure of 0.83%. The performance of the KNN was poor as it achieved 54% of accuracy, with precision of 0.64%, recall of 0.58% and F-Measure of 0.60%.



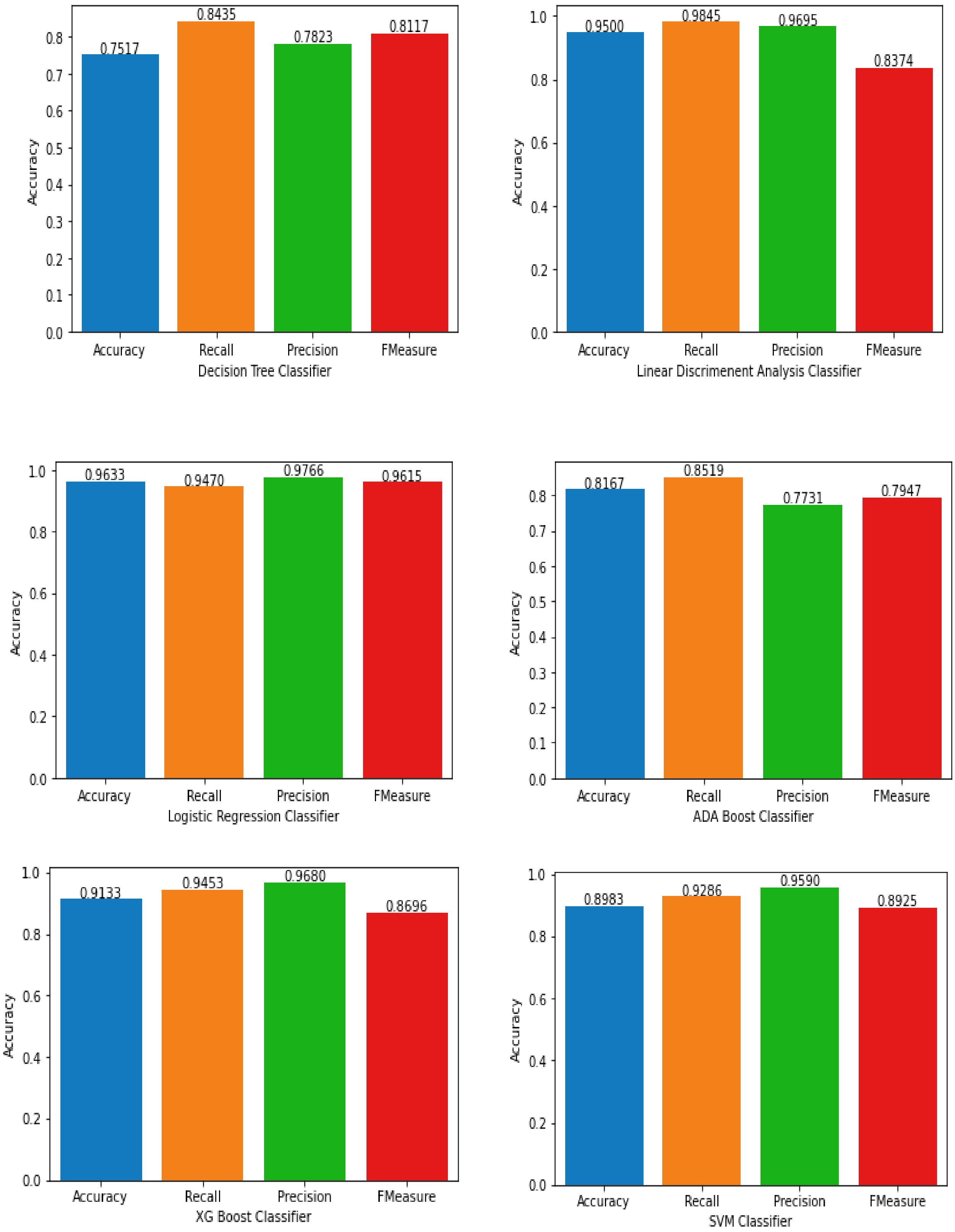
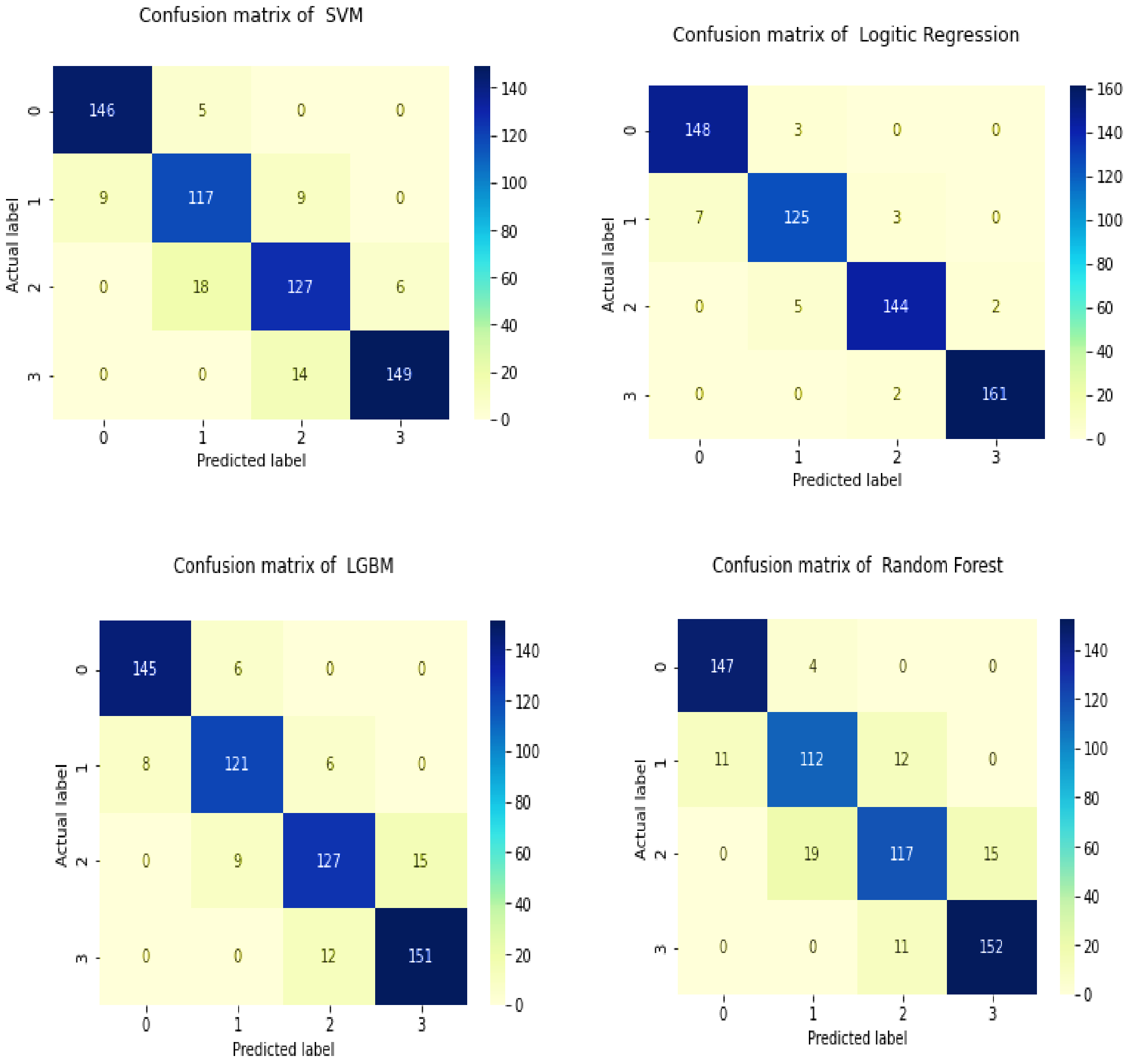


Figure 4. Results Evaluation of Used Models In Terms Of Accuracy, Precision, Recall And F-Measure

Additionally, NB demonstrated a contented 82% accuracy, with precision of 0.89%, recall of 0.90% and F-Measure of 0.90%. Similarly, DT achieved an accuracy of 75%, precision of 0.84%, recall of 0.78% and F-Measure of 0.81%. The AdaBoost has 81% of accuracy, precision of 0.77%, and recall of 0.85% and F-Measure of 0.79%, while the SVM achieved 88% of accuracy, precision of 0.95%, recall of 0.92% and F-Measure of 0.89%. The RF achieved 88% of accuracy, precision of 0.96%, recall of 0.91% as F-Measure of 0.93%, while the LR classifier performed 96% of accuracy, precision of 0.97%, recall of 0.94% and F-Measure of 0.96%. Next, the confusion matrixes achieved by all models are displayed in Figure 5.



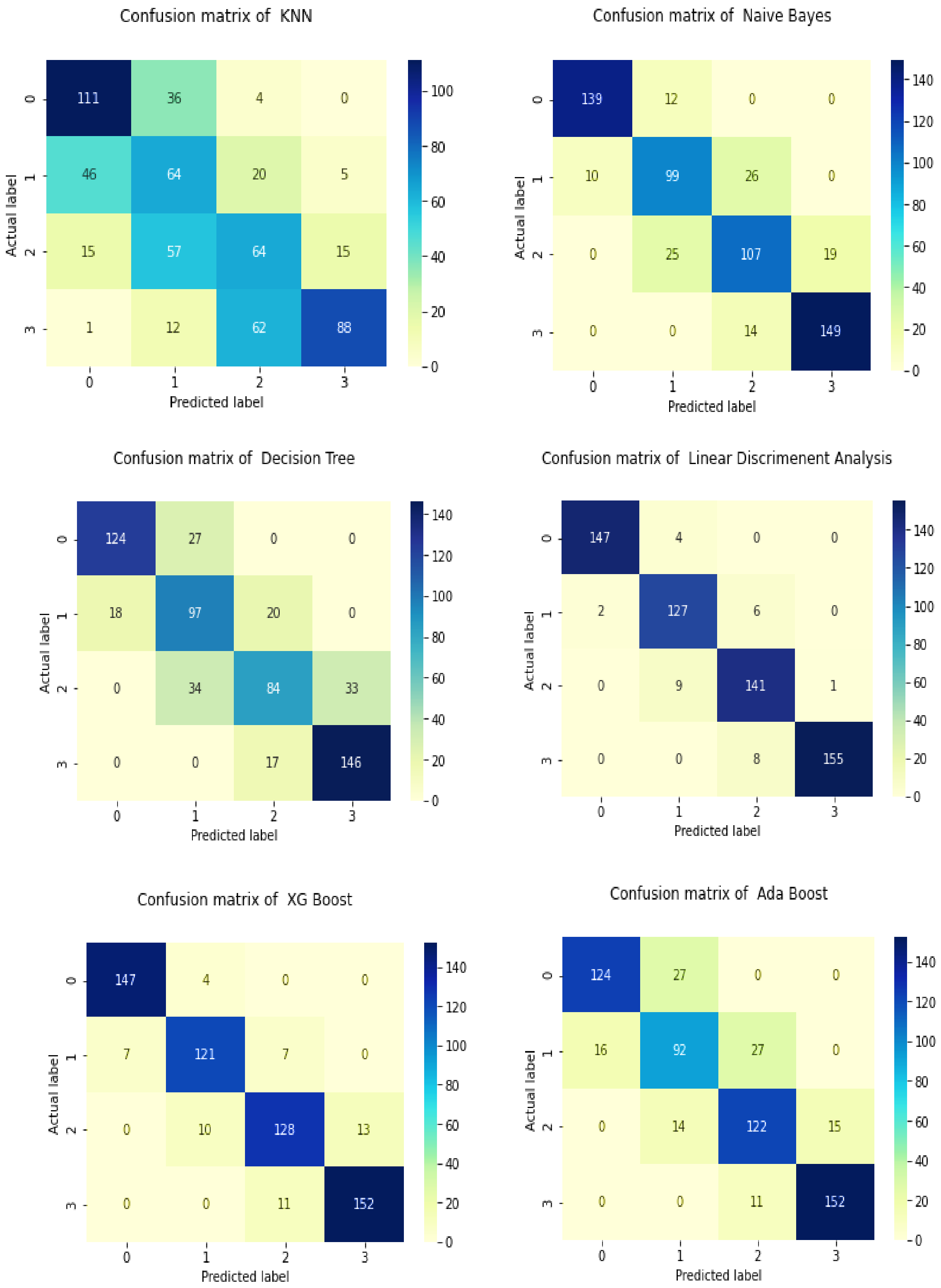


Figure 5. Confusion Matrix of Used Machine Learning Models For Mobile Price Prediction

Figure 5 shows the confusion matrix achieved by machine learning model KNN, NB, LDA, DT, LGB, RF , XGBoost, AdaBoost, SVM and LR model. Different machine learning models can achieve different types of performance (poor, good, excellent) for a single task. It is because of the architecture and learning nature of different machine learning models. It can happen with the type and nature of data used for analysis. In our proposed study, KNN achieved very poor results. It is because of the nature and learning paradigm of KNN models. As KNN works on local neighbourhoods, it is computationally expensive and less effective with large datasets. So, it causes poor results in this case. Similarly, DTs have less impact on large datasets because of its memory usage, lack of global optimization and overfitting, etc. However, this research has been conducted through feature engineering, which helps DT learn a bit more than KNN. The result of boosting algorithms, NB, SVM, LDA and RF was good because all the classifiers can deal with large and complex datasets compared to what was discussed above. Some of the key features of these algorithms that affect the proposed study results are the probability approach of NB, a combination of weak learners and the handling of complex relations by boosting algorithm, the effectiveness of high-dimensionality rates and robustness to overfitting by SVM and ensemble method with the importance of feature are the key features of RF.

6. CONCLUSION

Determining the mobile's actual pricing and estimating its position in the market is crucial for effective marketing and a product's successful launch. Different studies have been proposed for mobile cost prediction. However, for the said purpose, this study proposed ten famous machine learning classifiers like KNN, NB, XGBoost, DT, LGB, LDA, SVM, AdaBoost, LR and RF. The evaluation results indicated that for all the used machine learning in this paper, the LR classifier outperforms the other state-of-the-art models for mobile phone price prediction.

ACKNOWLEDGEMENT

The authors thank the College of Computing and Informatics, Saudi Electronic University, Riyadh, KSA and The University of Agriculture Peshawar for supporting this study.

AUTHOR CONTRIBUTIONS

Saima Anwar Lashari: Conceptualization, Data Curation,
Muhammad Muntazir Khan: Methodology, Validation, Writing – Original Draft Preparation;
Abdullah Khan: Project Administration, Supervision, Writing Review.
Sana Salahuddin: Writing – Review and Editing;
Muhammad Noman Atta: Writing – Review and Editing;

CONFLICT OF INTERESTS

The authors have no conflict of interests.

ETHICS STATEMENTS

Dataset used in the proposed study is collected from online repository Kaggle. The dataset is named as "Mobile_price_predictia" and link for dataset is https://www.kaggle.com/code/amankumar1007/mobile-price-predictiona/data.

REFERENCES

[1] Y. Chen, "Prediction of Different Types of Mobile Phone Prices based on Machine Learning Models," *Highlights in Science, Engineering and Technology*, vol. 92, pp. 275-279, 2024. doi: 10.54097/shgcew53.

[2] A. Saeed, A. Mukhtar, Y. Arafat, M. Abbas, and A. Saeed, "Intelligent Assessment of Secondhand Mobile Phone Prices by Machine Learning Techniques," *Journal of Computing and Biomedical Informatics (JCBI)*, 2024. https://jcbi.org/index.php/Main/article/view/351/262

[3] A. V. Kiran and R. Jebakumar, "Prediction of mobile phone price class using supervised machine learning techniques," *International Journal of Innovative Science, Research and Technology*, vol. 7, no. 1, pp. 248-251, 2022. doi: 10.5281/zenodo.5897944.

[4] M. Asim and Z. Khan, "Mobile price class prediction using machine learning techniques," *International Journal of Computer Applications*, vol. 179, no. 29, pp. 6-11, 2018. doi: 10.5120/ijca2018916555.

[5] K. S. Kalaivani, N. Priyadharshini, S. Nivedhashri, and R. Nandhini, “Predicting the price range of mobile phones using machine learning techniques,” in *AIP Conference Proceedings*, vol. 2387, no. 1, p. 140010, AIP Publishing, 2021, doi: 10.1063/5.0068605.

[6] W. Froelich and P. Hajek, “Combining rough set-based relevance and redundancy for the ranking and selection of nominal features,” *Procedia Computer Science*, vol. 176, pp. 1459-1468, 2020. doi: 10.1016/j.procs.2020.09.156.

[7] N. Hemageetha and G. M. Nasira, “Radial basis function model for vegetable price prediction,” in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pp. 424-428, 2013. doi: 10.1109/ICPRIME.2013.6496514.

[8] J. Zhou, Y. Qiu, S. Zhu, D. J. Armaghani, C. Li, H. Nguyen, and S. Yagiz, “Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate,” *Engineering Applications of Artificial Intelligence*, vol. 97, 104015, 2021. doi: 10.1016/j.engappai.2020.104015.

[9] R. Costache, A. Arabameri, H. Moayedi, Q. B. Pham, M. Santosh, H. Nguyen, et al., “Flash-flood potential index estimation using fuzzy logic combined with deep learning neural network, naïve Bayes, XGBoost and classification and regression tree,” *Geocarto International*, vol. 37, no. 23, pp. 6780-6807, 2022. doi: 10.1080/10106049.2021.1948109.

[10] H. A. Park, “An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain,” *Journal of Korean Academy of Nursing*, vol. 43, no. 2, pp. 154-164, 2013. doi: 10.4040/jkan.2013.43.2.154.

[11] S. R. Gunn, “Support vector machines for classification and regression,” *Technical Report*, vol. 14, no. 1, pp. 5-16, 1998. https://eprints.soton.ac.uk/256459/.

[12] M. Listiani, “Support vector regression analysis for price prediction in a car leasing application,” 2009. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=dc68aa0a77c71592e87e5d4097a2261c224184f8.

[13] M. Ali, F. Pervez, M.N. Atta, A. Khan, and A. Khan, “Sine Cosine Algorithm for Enhancing Convergence Rates of Artificial Neural Network: A Comparative Study”, *Journal of Engineering Technology and Applied Physics*, vol. 6, no. 2, pp. 34-37, 2024. doi: 10.33093/jetap.2024.6.1.3

[14] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” *Frontiers in Genetics*, vol. 9, 515, 2018. doi: 10.3389/fgene.2018.00515.

BIOGRAPHIES OF AUTHORS

	Saima Anwar Lashari was born on 1983, in Punjab Province, Pakistan. She enrolled in the University Tun Hussein Onn Malaysia to continue her PhD in 2013. During her Doctor of Philosophy (PhD) in Information Technology at University Tun Hussein Onn Malaysia (UTHM), she started her research journey under professional guidance of the supervision of Professor Dr. Rozati Ibraham. She is currently an Assistant Professor at the *College of Computing and Informatics, Saudi Electronic University, Riyadh, KSA*. She has published several research articles in the field of optimization and metaheuristics, neural network, data mining, prediction and Deep learning etc. Her main research interests include, knowledge-based systems, and data mining, optimization, predication, and web mining.
	Muhammad Muntazir Khan was born at Dargai distrislct Malakand KPK in 1992. He completed his MCS from. UNIVERSITY OF MALAKAND chakdara in 2015. Further he completed his MS CS from Agriculture University Peshawar in 2023 under the supervision of Dr. Abdullah assistant professor at AUP peshawar. His field of interest is Machine learning and deep learning. And published several publication to international journals in the field of machine learning and deep learning.
	Abdullah Khan was born on February 06 1985, in Dir (Lower), KPK Province, Pakistan. He did his BSc degree from Malakand University during 2004-2006. In 2006, he joined University of Science and Technology, Bannu, KPK Province, Pakistan for MSc in Computer Science. He later enrolled in the University Tun Hussein Onn Malaysia to continue his PhD at the end of 2010. During his Doctor of Philosophy (PhD) in Information Technology at University Tun Hussein Onn Malaysia (UTHM), he started his research journey under professional guidance of the supervision of Professor Dr. Nazri Mohd. Nawi. He is currently an Assistant Professor at the Institute of Computer Sciences and Information Technology faculty of Management and Computer Sciences the University of Agriculture, Peshawar, Pakistan. He has published several research articles in the field of optimization and metaheuristics, neural network, data mining, prediction and Deep learning etc. His main research interests include hybrid neural networks, knowledge-based systems, and data mining, deep learning, optimization, predication, and web mining.

	Sana Salahuddin was born in Peshawar KPK Province, Pakistan. She did her BSc degree from University Agriculture Peshawar during 2012-2016. In 2017, she joined University Agriculture Peshawar, KPK Province, Pakistan for MSc in Computer Science. She is currently working as visiting lecturer at the Institute of Computer Sciences and Information Technology faculty of Management and Computer Sciences the University of Agriculture, Peshawar, Pakistan. She has published several research articles in the field of optimization and metaheuristics, neural network, data mining, and prediction. Her main research interests include hybrid neural networks, knowledge-based systems, and data mining, predication, and networking.
	Muhammad Nouman Atta did his BSCS degree at the University of Agriculture, Peshawar, Pakistan during 2016-2020. In 2021, he joined the University of Agriculture, Peshawar, Pakistan MS in Computer Science. He started his research journey under the professional guidance and supervision of Assistant Professor Dr. Abdullah. He is currently enrolled in PhD at the Institute of Computer Sciences and Information Technology Faculty of Management and Computer Sciences the University of Agriculture, Peshawar, Pakistan. He has published several research articles in the field of optimization and metaheuristics, neural network, and Deep learning etc. His main research interests include hybrid neural networks, knowledge-based systems, and data mining, deep learning, optimization, predication and web mining.